Discrete Applied Mathematics 24 (1989) 187–196
North-Holland

ANTI-COMMUTATIVE LANGUAGES AND n-CODES*

M. ITO
Faculty of Science, Kyoto Sangyo University, Kyoto 603, Japan

H. JÜRGENSEN
Department of Computer Science, The University of Western Ontario, London, Ontario, Canada N6A 5B7

H.J. SHYR
Department of Mathematics, National Chung Hsing University, Taichung, Taiwan

G. THIERRIN
Department of Mathematics, The University of Western Ontario, London, Ontario, Canada N6A 5B7

Received 27 October 1988

1. Introduction

Languages derived from or related to codes have an important role in the study of the combinatorics of words. There are various mechanisms and tools to define and analyse codes. In particular, many classes of codes can be obtained as the classes of antichains with respect to certain partial orders on free monoids. As very simple examples we only mention the prefix codes, the hypercodes, and the block codes. Some details are provided in [2, 5, 6, 10].

In this paper we study a hierarchy

$$C(X) \subseteq \cdots \subseteq C_n(X) \subseteq C_{n-1}(X) \subseteq \cdots \subseteq C_2(X) \subseteq C_1(X)$$

of classes of languages over an alphabet X with $C(X)$ the class of codes over X. The languages in $C_n(X)$ are called n-codes; an n-code is a language each of whose n-element subset is a code. The original motivation for considering this hierarchy came from the analysis of $C_2(X)$ which had been shown to be the set of antichains with respect to a partial order derived from anti-commutativity [2]. However, the n-code property seems to be interesting in its own right.

We first prove that we are indeed dealing with a proper hierarchy. There is a major difference between languages in $C_2(X)$ and those in $C_n(X)$ for $n>2$. Whereas $C_2(X)$ is the class of antichains with respect to a certain partial order,

* This work was supported by the Natural Science and Engineering Research Council of Canada, Grants A7877 and A0243.

there is no binary relation ϱ such that $C_n(X)$ would be the set of ϱ-independent languages for $n>2$. In this way, $C_n(X)$ is similar to $C(X)$. It is known that there is no length-preserving binary relation nor any positive compatible partial order with the class of codes being their antichains [5, 9]. The latter statement can be extended to $C_2(X)$ as well.

The n-code hierarchy is "skew" with respect to the Chomsky hierarchy of languages; that is, for any given language classes $F \subseteq F'$ of the Chomsky hierarchy and for any n there are n-codes in $F \setminus F'$ which are not $(n-1)$-codes. For example, the Thue set T of square-free words over an alphabet X with $|X|>2$ is a non-algebraic type 1 language and also a 2-code, but not a 3-code.

It is obvious that as a consequence of the decidability of the code property also the n-code property can be decided for finite languages. We show that for rational languages the 2-code property is decidable. For $n>2$ the decidability question is open.

The decidability result for $C_2(X)$ is based on certain structural properties of 2-codes concerning primitive words. In particular, the 2-codes over X are subsets of cross sections of the equivalence relation on X^* defined by equality of roots. Using this fact, further insight into the structure of 2-codes can be gained.

This paper has the following sections – in addition to this introduction: In Section 2 we introduce notation and basic notions. Items not defined there or in the subsequent sections can be found in the books [1, 4, 8, 9] which we use as standard references. In Section 3 the hierarchy of n-codes is introduced and their properties concerning binary relations are proved. The role of primitive words is investigated in Section 4. Moreover, in Section 4, the n-code hierarchy is compared to the Chomsky hierarchy, and some decidability results are proved. Finally, Section 5 contains a few concluding remarks.

2. Notation and basic notions

An alphabet is a finite nonempty set. Let X be an alphabet. Then X^* denotes the free monoid generated by X, that is, the set of all words over X, including the empty word 1, and $X^+ = X^* \setminus 1$. For $w \in X^*$, by $|w|$ we denote the length of w.

A language over X is a set $L \subseteq X^*$. For any language L and any $n \in \mathbb{N}$ where $\mathbb{N} = \{0, 1, 2, \ldots\}$ let

$$L^{(n)} = \{w \mid \exists v \in L : v^n = w\}.$$

A word w is called *primitive* if $w = u^n$ implies $n = 1$. Let Q denote the set of all primitive words over X, where the alphabet X is understood. For $w \in X^+$ let $\sqrt{w}$ denote the unique word $u \in Q$ such that $w = u^n$ for some $n \in \mathbb{N}$.

Let ϱ be a binary relation on X^*. A language $L \subseteq X^*$ is said to be *ϱ-independent* or a *ϱ-antichain* if $u, v \in L$ and $u \varrho v$ implies $u = v$.

As standard reference for formal languages and acceptors we use [4]. In particular, we use the following notation for families of languages over an alphabet X:

- $\mathrm{Fin}(X)$ = finite languages,
- $\mathrm{Rat}(X)$ = rational (= regular = type-3) languages,
- $\mathrm{Alg}(X)$ = algebraic (= context-free = type-2) languages,
- $\mathrm{Cs}(X)$ = context-sensitive (= type-1) languages,
- $\mathrm{Rec}(X)$ = recursive languages,
- $\mathrm{RE}(X)$ = recursively enumerable (= type-0) languages,
- $P(X)$ = 2^{X^*} = general languages,
- $\mathrm{D0L}(X)$ = deterministic 0 Lindenmayer languages.

3. *n*-codes and binary relations

Let X be a finite alphabet, $|X| \geq 2$, and let $n \in \mathbb{N}$. A language L over X is said to be an *n-code* if $L \subseteq X^+$, L is nonempty, and every subset of L with at most n elements is a code. L is said to be *anti-commutative* if $L \subseteq X^+$ and $uv \neq vu$ for $u, v \in L$, $u \neq v$. Let $C(X)$, $C_n(X)$, and $A(X)$ denote the families of codes, n-codes, and anti-commutative languages over X, respectively.

Clearly

$$C(X) \subseteq \cdots \subseteq C_n(X) \subseteq C_{n-1}(X) \subseteq \cdots \subseteq C_2(X) \subseteq C_1(X),$$

where $C_1(X)$ is trivial, that is,

$$C_1(X) = (2^{X^+} \setminus \{\emptyset\}).$$

Furthermore,

$$C_2(X) = A(X)$$

from the fact that a set $\{u, v\}$ is a code if and only if $uv \neq vu$ (see e.g. [9]). Clearly also

$$C(X) = \bigcap_{i=1}^{\infty} C_i(X).$$

That the above inclusions are proper is easily seen by the following example from [9]: Let

$$C = \{a_1, \ldots, a_n\} \subseteq X^+$$

be a code over X with $|C| = n$. The existence of C is guaranteed by the fact that any finitely generated free monoid can be embedded into X^+ when $|X| \geq 2$. Now consider

$$L = \{a_1, \ldots, a_n, a_1 \ldots a_n\}.$$

Obviously, L is an n-code but not an $(n+1)$-code.

In fact this example is just a special case of a more general construction.

Proposition 3.1. *Let C be any k-code or a code over X, and let $1<n\leq|C|$ and $2n-1\leq k$. If $a_1,\ldots,a_n$ are any n distinct elements of C, then the set*

$$L=C\cup\{a_1\ldots a_n\}$$

is an n-code but not an $(n+1)$-code.

Proof. It is immediate that L is not an $(n+1)$-code. In order to show that it is an n-code consider a set

$$L'=\{w_1,\ldots,w_{n-1}\}\cup\{a_1\ldots a_n\},$$

where $w_1,\ldots,w_{n-1}$ are $n-1$ distinct elements of C. Consider also the set

$$L''=\{w_1,\ldots,w_{n-1}\}\cup\{a_1,\ldots,a_n\}.$$

Since $L''\subseteq C$ and C is a k-code or a code, also L'' is a code as $k\geq 2n-1$.

Suppose L' is not a code. Thus there exists a word with two different representations over L',

$$x_1\ldots x_r=y_1\ldots y_m$$

with $x_i,y_j\in L'$ for $i=1,\ldots,r$ and $j=1,\ldots,m$.

Now, if $x_1=w_k$, $y_1=w_h$ for some k,h, then $x_1=y_1$ as L'' is a code. Therefore, we may assume that

$$x_1=w_1 \quad\text{and}\quad y_1=a_1\ldots a_n,$$

that is,

$$w_1x_2\ldots x_r=a_1a_2\ldots a_ny_2\ldots y_m.$$

Viewing the word over L'' yields the factorizations

$$w_1(x_{21}\ldots x_{2k_2})\ldots(x_{r1}\ldots x_{rk_r})=a_1\ldots a_n(y_{21}\ldots y_{2h_2})\ldots(y_{m1}\ldots y_{mh_m}),$$

where

$$x_i=(x_{i1}\ldots x_{ik_i}) \quad\text{and}\quad y_j=(y_{j1}\ldots y_{jh_j}).$$

Therefore, $w_1=a_1$ and

$$(x_{21}\ldots x_{2k_2})\ldots(x_{r1}\ldots x_{rk_r})=a_2\ldots a_n(y_{21}\ldots y_{2h_2})\ldots(y_{m1}\ldots y_{mh_m}).$$

If

$$x_2=x_{21}\ldots x_{2k_2}=a_1\ldots a_n,$$

then $a_1=a_2$, a contradiction! Therefore, $x_2=w_s$ for some s and $w_s=a_2$. Iterating this argument yields $x_i=a_i$ for $i=1,\ldots,n$ and also $x_i\in\{w_1,\ldots,w_{n-1}\}$. This is impossible as all words $a_1,\ldots,a_n$ were chosen distinct. □

In Proposition 3.1 $k=n$ is not possible. For the set

$$C=a^+b^+\cup b^+a^+\cup\{aba^2b^2a^3b^3\}$$

in a 3-code while

$$C \cup \{a^2b^2a^3b^3ab\}$$

is not a 3-code.

The set

$$B = a^+b^+ \cup b^+a^+$$

is an example of a language in $C_3(X) \setminus C_4(X)$ which is not obtained by the construction given in Proposition 3.1. So far no generalization of this example to arbitrary n is known.

The family $C_2(X)$ of 2-codes is of particular interest. It was proved in [2] that $C_2(X)$ coincides with the family of antichains with respect to the partial order $\leq_c$ on X^* defined by

$$x \leq_c y \quad \Leftrightarrow \quad \exists u \in X^*: \; y = xu = ux.$$

For further results concerning the relation between partial orders on X^* and codes the reader is referred to [2, 5, 9, 10].

A binary relation $\varrho \subseteq X^* \times X^*$ is called *length-preserving* if it satisfies the following conditions:

(1) $\forall u \in X^*: u \varrho u$;
(2) $u \varrho v$ implies $|u| \leq |v|$;
(3) $u \varrho v$ and $|u| = |v|$ together imply $u = v$.

A length-preserving binary relation on X^* is reflexive and anti-symmetric, but not necessarily transitive or compatible. A binary relation $\varrho \subseteq X^* \times X^*$ is said to be *positive* if

(4) $\forall u \in X^*: 1 \varrho u$

holds true. Observe that the partial order $\leq_c$ is both length-preserving and positive, but not compatible. Positive length-preserving partial orders are called *strict* in [9] and elsewhere.

Whereas quite a few interesting classes of codes – the prefix codes, suffix codes, bi-prefix codes, hypercodes, to mention only a few examples – can be characterized as the classes of antichains of certain partial orders on X^*, the class $C(X)$ of codes cannot be described in such a way: there is no length-preserving binary relation nor any positive compatible partial order on X^*, say ϱ, such that $C(X)$ coincides with the class of ϱ-antichains [5, 9]. A far stronger statement can be proved for the classes $C_n(X)$.

Proposition 3.2. *Let $n \geq 3$ and $|X| \geq 2$. There is no binary relation ϱ on X^* such that $C_n(X)$ is the class of all ϱ-antichains.*

For $n = 2$ a slightly weaker statement can be made.

Proposition 3.3. *Let $|X| \geq 2$. There is no compatible binary relation ϱ over X^* such that $C_2(X)$ is the class of all ϱ-antichains.*

These results seem to indicate that n-codes are rather complex objects. This will be clarified to a certain extent in the sequel.

4. n-codes and primitive words

It is well known that a pair $x, y \in X^+$ of words forms a code if and only if $xy \neq yx$ or, equivalently, if $\sqrt{x} \neq \sqrt{y}$ (see [9], for example). For words $x, y \in X^+$ define the relation $\sim_1$ by

$$x \sim_1 y \quad \Leftrightarrow \quad \sqrt{x} = \sqrt{y}.$$

Obviously, $\sim_1$ is an equivalence relation on X^+.

Lemma 4.1. *Let $|X| \geq 2$ and $L \subseteq X^+$. The following statements are equivalent:*

(1) $L \in A(X)$;
(2) $L \in C_2(X)$;
(3) *L is contained in a cross section of $\sim_1$.*

L is a maximal 2-code if and only if it is a cross section of $\sim_1$.

For a proof of Lemma 4.1 see [2]. Its assertion (3) allows for a useful 1-1-correspondence between 2-codes L over X and mappings

$$f_L : Q \to \mathbb{N},$$

which is given by

$$u \in L \quad \Leftrightarrow \quad f(\sqrt{u}) \neq 0 \wedge \sqrt{u}^{f(\sqrt{u})} = u.$$

Conversely if f is a mapping of Q into $\mathbb{N}$, then

$$L_f = \{u^{f(u)} \mid u \in Q \wedge f(u) \neq 0\}.$$

This representation of $C_2(X)$ implies the following corollary:

Corollary 4.2. *For $|X| \geq 2$ one has $|C_2(X)| = \aleph_1$, and thus $C_2(X)$ is not recursively enumerable. In particular there are 2-codes which are not even type 0 languages.*

This result seems to indicate that the classes of n-codes may be "skew" with respect to standard language classes. Further details substantiating this impression will be provided in a follow-up paper.

Proposition 4.3. *Let $|X| \geq 2$ and let $L \in C_2(X)$. The following properties obtain:*

(1) if L is rational, *then* f_L *is bounded*;
(2) *if* f_L *is unbounded, then the order of the elements of the syntactic monoid* syn L *of* L *is unbounded*;
(3) *there is a context-free 2-code* L *with* f_L *unbounded.*

Proof. As f_L is unbounded, for any $k \in \mathbb{N}$ there exists $m > k$ such that $f^m \in L$ for some $f \in Q$, and therefore, $f^n \notin L$ for $n \neq m$. Thus, the words $f, f^2, \ldots, f^m$ are pairwise incongruent modulo the principal congruence P_L of L. This implies (2), and therefore syn L is infinite, that is, L is not rational. Now consider the language

$$L = \bigcup_{i=0}^{\infty} ab^i(ab^+)^i$$

over the alphabet $X = \{a, b\}$. Obviously, L is context-free. To show that L is a 2-code suppose that

$$ab^i ab^{h_1} ab^{h_2} \ldots ab^{h_r} = f^s, \qquad ab^j ab^{k_1} ab^{k_2} \ldots ab^{k_t} = f^t$$

for some $f \in Q$ and $s, t \geq 1$.

If $f = ab^i$, then $f = ab^j$, that is, $i = j$ and therefore $s = t$.

Otherwise, we have to assume that

$$f = ab^i ab^{h_1} \ldots ab^{h_p} = ab^j ab^{k_1} \ldots ab^{k_q}$$

for some $p, q \geq 1$. Then obviously $i = j$ and $p = q$ and thus $s = t$.

Therefore, L is a 2-code, in fact, it is a code. As $(ab^i)^{i+1} \in L$ and $ab^i \in Q$ for every $i \geq 1$ it follows that f_L is unbounded. □

Observe that boundedness of f_L does not imply rationality for L. The language $L = Q$ is such an example of a nonrational language with f_L bounded.

For a language L over X let $\sqrt{L}$ denote the language

$$\sqrt{L} = \{u \mid u \in Q \wedge \exists v \in L : u = \sqrt{v}\}.$$

If $L \neq \emptyset$, then $\sqrt{L} \in C_2(X)$. For $L \in C_2(X)$ one has $\sqrt{L} = Q$ if and only if L is a maximal 2-code. However, the following result implies that $\sqrt{L} \neq Q$ if $L \in C_n(X)$ for $n \geq 3$.

Proposition 4.4. *Let* $|X| \geq 2$ *and* $n \geq 3$. *For every n-code* $L \subseteq X^*$ *the set* $Q \setminus \sqrt{L}$ *is infinite.*

Corollary 4.5. *Let* $|X| \geq 2$. *Then the following statements hold true:*

(1) *if* $L \in C_2(X) \cap \mathrm{Rat}(X)$ *and* L *is infinite, then* $L \cap Q$ *is infinite*;
(2) *if* $L \in C_2(X) \cap \mathrm{Rat}(X)$ *and* L *is infinite, then* $L \cap Q$ *is rational if and only if* $L \cap (X^* \setminus Q)$ *is finite*;
(3) *for any finite set* $M \neq \emptyset$, $M \subseteq \{1, 2, \ldots\}$, *there is a rational 2-code* L *such that* $L \cap Q^{(m)}$ *is infinite for all* $m \in M$.

Corollary 4.6. *Let $|X| \geq 2$. Every infinite rational code over X contains infinitely many primitive words.*

We conclude this section with a description of the relation between D0L languages and 2-codes.

Proposition 4.7. *Let $|X| \geq 2$ and let $L \in \mathrm{D0L}(X)$ be infinite. If $L \notin C_2(X)$, then there is an integer k such that $|L'| \leq k$ for every $L' \subseteq L$ which is a 2-code.*

Proof. Let L be an infinite D0L language generated by the D0L system $G = (X, h, w_0)$, and let $w_i = h^i(w_0)$ for $i \in \mathbb{N}$. Suppose that L is not a 2-code. Then there exist $i, k \in \mathbb{N}$, $i < k$, such that $\{w_i, w_k\}$ is not a code, that is, $w_i = p^n$, $w_k = p^m$ for some $p \in Q$ and $n, m \geq 1$, $n \neq m$. Choose k minimal with this property. Then the set

$$\bar{L} = \{w_0, w_1, \ldots, w_{i-1}\}$$

is a 2-code.

Now let $t = k - i$. From $w_i = p^n$ and $w_k = w_{i+t} = p^m = h^i(p^n) = (h^i(p))^n$ it follows that $h^i(p) = p^l$ for some $l \geq 1$. Therefore, $m = ln$ and

$$w_{i+rt+s} = h^{rt}(h^s(w_i)) = h^{rt}(h^s(p))^n = (h^s(p))^{l^{rt}n}$$

for $r \in \mathbb{N}$ and $s = 0, 1, \ldots, t-1$. Let

$$L_s = \{w_{i+rt+s} \mid r \in \mathbb{N}\}.$$

Then

$$L = \bar{L} \cup \bigcup_{s=0}^{t-1} L_s$$

with L_s infinite for all s. If L' is any 2-code contained in L, then $|L' \cap L_s| \leq 1$ and, therefore, $|L'| \leq t + i = k$. □

Corollary 4.8. *Every infinite context-free* D0L *language is a 2-code.*

Proof. This result follows from the preceding proof by the "pumping lemma" for context-free languages. It can also be obtained as a weak version of a result due to [3] which states that every infinite context-free D0L language is a prefix code or a suffix code, which implies that it is a 2-code. □

As an example of a D0L language which is a 2-code but not a 3-code consider the set

$$\{a, b, ab, bab, abbab, \ldots\},$$

that is, the *Fibonacci language* over $X = \{a, b\}$ which is generated by the D0L rules $a \to b$, $b \to ab$.

We now proceed to prove that the property of being a 2-code is decidable for

rational languages. The following statement provides the main argument for the proof.

Proposition 4.9. *Let $L \in \mathrm{Rat}(X)$. It is decidable whether there exists a word $w \in X^+$ such that $w^i \in L$ and $w^j \in L$ for two different powers of w.*

Proof. Let $A = (X, S, \delta, q_0, F)$ be a finite state acceptor with $L = L(A)$. For $q_\alpha, q_\beta \in S$ let

$$A_{q_\alpha q_\beta} = (X, S, \delta, q_\alpha, \{q_\beta\}),$$

and let

$$L_{q_\alpha q_\beta} = L(A_{q_\alpha q_\beta}).$$

Obviously, the following two statements are equivalent:

(1) $\exists w \in X^+ \ \exists i > j > 0$: $w^i, w^j \in L$, and

(2) $\exists q_1, q_2, \ldots \in S$:

$$\bigcap_{i \geq 1} L_{q_{i-1} q_i} \setminus \{1\} \neq \emptyset, \quad \text{and} \quad |\{q_1, q_2, \ldots\} \cap F| \geq 2.$$

Thus, in order to decide (1) one could try to decide (2).

Observe first that the sequence of states reached by consecutive powers of a fixed input word is ultimately periodic. Thus, if

$$q_i = \delta(q_0, w^i),$$

then the sequence has the form

$$q_0, q_1, \ldots, q_i, q_{i+1}, \ldots, q_{i+p} = q_i.$$

Therefore in deciding (2) we may restrict ourselves to considering sequences of this form where $i + p \leq n - 1$ with $n = |S|$. There are no more than $n! 2^n$ such sequences and one checks each of them separately.

Now consider such a sequence. The condition

$$|\{q_1, q_2, \ldots\} \cap F| \geq 2$$

is satisfied if and only if

(a) there are two indices $0 \leq \alpha < \beta < i$ with $q_\alpha, q_\beta \in F$, or

(b) there is $i \leq \alpha < i + p$ with $q_\alpha \in F$.

Thus the above condition can easily be checked. Finally,

$$\bigcap_{i=1}^{\infty} L_{q_{i-1} q_i} = \bigcap_{i=1}^{i+p} L_{q_{i-1} q_i},$$

and therefore also the condition

$$\bigcap_{i=1}^{\infty} L_{q_{i-1} q_i} \setminus \{1\} \neq \emptyset$$

is decidable. □

Proposition 4.10. *Let* $L \in \mathrm{Rat}(X)$. *It is decidable whether* $L \in C_2(X)$ *holds. If* $L \in \mathrm{Fin}(X)$, *then it is also decidable whether* $L \in C_n(X)$ *for* $n > 2$.

Proof. The first statement is a consequence of Proposition 4.9. The second one follows from the fact that for testing the n-code property on a finite set it is sufficient to check the code property on its (finitely many) subsets of size n. □

5. Concluding remarks

This paper focusses on the following problem areas concerning n-codes:

(1) n-code hierarchy;
(2) definition by binary relations;
(3) relation to the set Q of primitive words;
(4) comparison with the Chomsky hierarchy;
(5) decidability of the n-code property.

Several results concerning more detailed structural descriptions of n-codes, their syntactic monoids, and properties like maximality have been omitted here and will be presented in a follow-up paper. Open problems abound – we mentioned the decidability or undecidability of the n-code property for rational languages; a more precise comparison to other language classes would be another simple example; some other more intricate ones have been omitted to keep the presentation consice.

References

[1] J. Berstel and D. Perrin, Theory of Codes (Academic Press, New York, 1985).
[2] P.H. Day and H.J. Shyr, Languages defined by some partial orders, Soochow J. Math. 9 (1983) 53–62.
[3] T. Head and G. Thierrin, Polynomially bounded D0L systems yield codes, in: L. Cummings, ed., Combinatorics on Words (Academic Press, New York, 1983) 167–174.
[4] J.E. Hopcroft and F.D. Ullman, Introduction to Automata Theory, Languages, and Computation (Addison-Wesley, Reading, MA, 1979).
[5] H. Jürgensen, H.J. Shyr and G. Thierrin, Codes and compatible partial orders on free monoids, Astérisque, to appear.
[6] H. Jürgensen and G. Thierrin, Infix codes, in: Proceedings Comp. Sci. Conf., Győr (1985).
[7] J. Karhumäki, On three-element codes, in: J. Paredaens, ed., Proceedings ICALP 1984, Lecture Notes Computer Science 172 (Springer, Berlin, 1984) 292–302.
[8] M. Lothaire, Combinatorics on Words (Addison-Wesley, Reading, MA, 1983).
[9] H.J. Shyr, Free monoids and languages, Lecture Notes, Dept. Math., Soochow Univ., 1979.
[10] H.J. Shyr and G. Thierrin, Codes and binary relations, in: Séminaire d'Algèbre P. Dubreil, 1975/76, Lecture Notes in Mathematics 586 (Springer, Berlin, 1977) 180–188.